分割表の列挙とランダム生成 (グレブナ-基底の理論的有効性と実践的有効性)

AUTHOR(S):

松井, 泰子

CITATION:

松井, 泰子. 分割表の列挙とランダム生成 (グレブナ-基底の理論的有効性と実践的有効性). 数理解析研究所講究録 2002, 1289: 94-109

ISSUE DATE:

2002-09

URL:

http://hdl.handle.net/2433/42508

数理解析研究所講究録 1289 巻 2002 年 94-109

分割表の列挙とランダム生成

東海大学・理学部・情報数理学科　松井　泰子 (Yasuko Matsui)

Department of Mathematical Sciences,

Tokai University

1 Introduction

分割表 (Contingency Tables) は統計データの分類に用いられ，統計データの検定のためには，周辺和を固定した分割表の列挙やサンプリングが必須である．

本稿では，第 2 章で分割表が用いられる背景を簡単に説明し，第 3 章で分割表と整数計画問題の関連について述べる．整数計画問題は，トーリックイデアルのグレブナー基底を用いて解ける [33] が，整数計画問題と周辺和を固定した分割表は密接な関係があるため，その列挙やサンプリングは，グレブナー基底を用いて実行出来る事が知られているので紹介する．第 4 章では，周辺和を固定した分割表の counting と列挙に関する計算複雑さについて示す．第 5 章で，周辺和を固定した分割表をサンプリングするためのマルコフ連鎖の構成法を紹介する．第 6 章では，マルコフ連鎖の構成に関する最近の結果を紹介する．

2 統計の背景

本章では，分割表を定義し，分割表が用いられる統計の背景を説明する．

統計データは、複数の属性で分類される．第 i_1 の属性で分類した後，分類した各々のデータを第 i_2 の属性で分類し，第 i_k の属性で分類してまとめたものを k 元分割表 (k-way contingency table) とよぶ．第 j の属性によって m_j 個の類が定義されている場合，k 元分割表は $m_1 \times m_2 \times \cdots \times m_k$ 分割表，$i_1 \times i_2 \times \cdots i_k$ 分割表とも呼ばれる．分割表では，各属性間の相関の有無が検定される．以下に 2 元分割表を例に，検定を説明する．

いま，N 個のデータに対し 2 つの属性 A, B があり，属性 A は,$A_1, A_2, \cdots, A_d$ に，属性 B は,$B_1, B_2, \cdots, B_n$ に分割されているとする．データが A_i, B_j に属する確率を各々 $p_{i\cdot}, p_{\cdot j}$，データが A_i, B_j に共に属する確率を p_{ij}，$|A_i| = x_{i\cdot}, |B_j| = x_{\cdot j}, |A_i \cap B_j| = x_{ij}$ とし，総度数

を N と表す．このとき，2つの属性 A, B が独立であるという帰無仮説 $H : p_{ij} = p_{i\cdot}p_{\cdot j}$ を検定して属性間の相関を調べる．表1に2元分割表の例を示す．ただし，$I = \{1, 2, \ldots, d\}, J = \{1, 2, \ldots, n\}$ とする．

表 1:　2元分割表 ($d \times n$ 分割表,$I \times J$ 分割表)

$A \setminus B$	B_1	B_2	$\cdots$	B_n	計
A_1	x_{11}	x_{12}	$\cdots$	x_{1n}	$x_{1\cdot}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2n}	$x_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_d	x_{d1}	x_{d2}	$\cdots$	x_{dn}	$x_{d\cdot}$
計	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot n}$	N

帰無仮説の検定方法の1つに χ^2 適合度検定がある．実際に観測されたセルのデータ $x = (x_{ij})$ を帰無仮説のもとで統計量

$$\chi^2(x) = \sum_{i=1}^{d}\sum_{j=1}^{n} \frac{x_{ij} - (\frac{x_{i\cdot}x_{\cdot j}}{n})^2}{(\frac{x_{i\cdot}x_{\cdot j}}{n})}$$

を用いて検定するものである．帰無仮説 H が正しいならば，$\chi^2(x)$ は $n \to \infty$ で漸近的に自由度 $(d-1)(n-1)$ の χ^2 分布に従うことが知られている．

しかし，総度数 N が十分に大きくない場合には漸近分布論の当てはまりの悪さが問題となる．そのときは，帰無仮説のもとでの x の分布は多項超幾何分布 (*multiple hypergeometric distribution*)$H(x)$ に従うと仮定する．$H(x)$ は，行和，列和 (以後，周辺和とする) が $(x_{1\cdot}, \ldots, x_{d\cdot}), (x_{\cdot 1}, \ldots, x_{\cdot n})$ に一致するような分割表 x に対してのみ用いられ，定義は以下のとおりである．

$$H(x) = Pr(X = x) = \frac{(\Pi_{i=1}^{d} x_{i\cdot}!)(\Pi_{j=1}^{n} x_{\cdot j}!)}{N!\Pi_{i=1}^{d}\Pi_{j=1}^{n} x_{ij}!}$$

ここで，周辺和を固定した分割表の集合を Ω，観測された分割表を x^* とし，

$$p = \sum_{x \in \Omega, \chi^2(x) \geq \chi^2(x^*)} H(x)$$

で定義される有意確率を p 値と呼ぶ．p 値と適当な有意水準 α とを比較することで，χ^2 検定が行えるが，p 値の厳密計算は $|\Omega|$ が大きいと困難となる．例えば，4×4 分割表で，行和，列和が各々$(1000, 9000, 3000, 440), (2000, 4000, 200, 7240)$ となる分割表は 10^{23} 個程度存在するので，分割表のサイズが大きくなると列挙も容易では無い [36]．そこで，帰

無仮説のもとで $x \in \Omega$ をサンプリングし，p 値を計算する方法が提案されている．これを**マルコフ連鎖モンテカルロ法** (Markov Chain Monte Carlo Method;MCMC 法) という．MCMC 法では，仮定した分布を極限分布としてもつマルコフ連鎖において十分反復した後サンプリングを行う．マルコフ連鎖は，定常分布が仮定した分布 (多項超幾何分布，一様分布 (*uniform distribution*) 等) に一致するような既約かつ非周期的なものを構成し，任意の初期状態から，この連鎖上で十分な回数推移を行ってサンプリングを行うものである．MCMC 法は統計物理の分野で提案され，1990 年以降研究が進み現在では幅広く用いられている．

3 分割表と整数計画問題

本章では，分割表と整数計画問題の関係について述べる．いま $I \times J$ 分割表について考える．ここで $|I| = d, |J| = n$ とし，各セルを x_{ij}，行和と列和を各々 a_i, b_j とすると，$\sum_{j=1}^{n} x_{ij} = a_i, i \in I$, $\sum_{i=1}^{d} x_{ij} = b_j, j \in J$ と表すことができる．

整数計画問題の一つに**輸送問題** (*Transportation problem*) がある．輸送問題は，n 箇所の供給地 (工場等)I，m 箇所の需要地 (消費地等)J，供給量 $a_i, i \in I$，需要量 $b_j, j \in J$ と供給地と需要地間の単位当りの輸送コスト $c_{ij}, i \in I, j \in J$ が与えられたときに，輸送コストを最小とするような供給地から需要地へのフロー (x_{ij}) を求める問題 (2 部グラフ上のフロー問題) であり，以下のように定式化できる．

輸送問題

$$
\begin{aligned}
&\text{Minimize.} && \sum_{i=1}^{d} \sum_{j=1}^{n} c_{ij} x_{ij} \\
&\text{Subject to.} && \sum_{j=1}^{n} x_{ij} = a_i,\ i = 1, \ldots, d, \\
& && \sum_{i=1}^{d} x_{ij} = b_j,\ j = 1, \ldots, n, \\
& && x_{ij} \in R_{\geq 0}.
\end{aligned}
$$

輸送問題は整数最適解をもつので，最後の制約式を $x_{ij} \in Z_{\geq 0}$ としても良い．

定義 1 ([32]) 行列 A の各小行列式が $\{0, \pm 1\}$ であるとき，A は *totally unimodular* であるという．

与えられた $\{-1, 0, 1\}-$ 行列が totally unimodular であるか否かの判定は，$\mathcal{NP}$-complete である [21]．しかし，輸送問題の制約式の係数行列である $\{0, 1\}-$ 行列については，totally unimodular となることが知られている [32]．

定理 1 (Schrijver [32]) *A を totally unimodular 行列，b を整数ベクトルとしたとき，多面体 $P := \{x | Ax \leq b\}$ は P 中の整数ベクトルの凸包である．* □

輸送問題の制約式の係数行列は totally unimodular 行列なので，輸送問題多面体の頂点は整数格子点である．また，周辺和 a_i, b_j の 2 元分割表 (x_{ij}) と，供給量 a_i と需要量 b_j の輸送問題の実行可能 (整数) 解 x_{ij} が一対一対応しているので，周辺和を固定した 2 元分割表の counting は，輸送問題多面体中の整数格子点の counting に等しい．したがって，周辺和を固定した 2 元分割表の counting が出来れば，Ehrhart 多項式 [24] を用いて輸送問題多面体の volume の計算が可能である．

特殊な 3 元分割表に関しては，以下が成立する事が示されている．

系 1 (Sturmfels [33]) *$2 \times J \times K$ 分割表の係数行列は unimodular である．* □

しかし，$\mathbf{3} \times J \times K$ 分割表の係数行列は unimodular ではない [15] ので，前出の方法による輸送問題多面体の volume 計算は，一般次元には拡張出来ない．

4 分割表の counting と列挙

前章で述べたように，分割表の列挙に関しては多面体の幾何学が深く関わっている．本章では，多面体中の整数格子点の counting 問題等の計算の複雑さに関する既往の研究を紹介しながら，周辺和を固定した分割表の counting や列挙について述べる．

周辺和を固定した分割表の counting の計算複雑さに関しては，次のような結果が得られている．

定理 2 (Dyer, Kannan & Mount [19]) *周辺和を固定した分割表の counting は，$2 \times n$ 分割表の場合でさえ $\sharp$P-complete である．* □

証明では，係数が正整数である線形不等式系で表された多面体の volume 計算が $\sharp$P-hard である [16] 事を用いて，輸送問題多面体の volume 計算 ($= 2 \times n$ 分割表の couting) の計算の複雑さを示している．

上の定理では，分割表の counting は効率良く行えない事を示したが，周辺和を固定した 2 元分割表は，容易に求める事ができる．なぜなら輸送問題は，問題の入力サイズの多項式時間のオーダーで解を得ることが可能だからである．

しかし，周辺和を固定した 3 元以上の分割表の存在性に関しては，以下の定理が証明さ

定理 3 (Irving & Jerrum [25]) 与えられた周辺和を満たす3元分割表の存在性の判定は，*NP-complete* である．□

3元分割表に関する問題の計算の複雑さに関しては，De Loera & Onn [10] が詳しい．

De Loera & Sturmfels [11] は，多面体の volume 計算は，Chamber 多項式を用いて計算できる事を示した．さらに，代数幾何的アプローチのアルゴリズムとして，ある射影多様体の (1) コホモロジー環の構成アルゴリズム，と (2)Todd クラスの積分計算アルゴリズムを組合せたものを提案した．(1),(2) のアルゴリズムは共にグレブナー基底の計算によって実行できることが示されている．

Sturmfels [33] は，Avis & 福田の逆探索法 [6] の枠組みを利用すれば，グレブナー基底を用いて周辺和を固定した2元分割表が列挙できる事を示している．

5 分割表のサンプリング

周辺和を固定した分割表のサンプリングは，状態空間を分割表全体とし，仮定した分布に収束するような既約かつ非周期的なマルコフ連鎖を構築して行う．以下に，マルコフ連鎖に関する用語の定義を簡単にする．

任意の $n, j \in Z_{\geq 0}$ に対し，確率過程 $\{X_n; n = 0, 1, 2, \ldots\}$ が条件:

$$\Pr\{X_{n+1} = j | X_0, X_1, \ldots, X_n\} = \Pr\{X_{n+1} = j | X_n\}$$

を満たすとき，その確率過程をマルコフ連鎖 $\mathcal{M}$ という．状態 i から状態 j への推移確率を $P(i,j)$ と表し，$P(i,j)$ を i,j 要素とする行列 $\boldsymbol{P}$ を推移確率行列という．推移確率行列 $\boldsymbol{P}$ のマルコフ連鎖 $\mathcal{M}$ が時刻 t に状態 i にある確率を $\pi_n(i)$ とし，$\boldsymbol{\pi}_n = (\pi_n(1), \pi_n(2), \ldots)$ とすると，$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1}\boldsymbol{P}$ $(n \geq 1)$ が成り立ち，$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0\boldsymbol{P}^n$ が得られる．π_n が $n \to \infty$ で $\boldsymbol{\pi}_0$ と無関係な極限分布に近づくとき，それを $\boldsymbol{\pi}$ とすると，$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$ が成り立ち，この式を平衡方程式と呼ぶ．平衡方程式を満たす確率分布 $\boldsymbol{\pi}$ をマルコフ連鎖 $\mathcal{M}$ の定常分布と呼ぶ．

分割表のサンプリングにマルコフ連鎖を初めて導入したのは，Diaconis & Saloff-Coste である [14]．彼らは提案した (一様分布に収束する)$I \times J$ 分割表をサンプリングするためのマルコフ連鎖が，入力サイズの擬多項式時間で収束することを証明した (ただし，行数と列数は定数とみなしている)．推移方法は以下のとおりである．

Diaconis & Saloff-Coste のマルコフ連鎖の推移方法

Step 1. $I \times J$ 分割表 x 中の任意の相異なる 2 行 $i_1, i_2 \in I(i_1 < i_2)$，相異なる 2 列 $j_1 j_2 \in J(j_1 < j_2)$ を選ぶ.

Step 2. x の $\begin{bmatrix} (i_1,j_1) & (i_1,j_2) \\ (i_2,j_1) & (i_2,j_2) \end{bmatrix}$ 要素に $\begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}$ もしくは $\begin{bmatrix} -1 & +1 \\ +1 & -1 \end{bmatrix}$ を 1/2 の確率で加えて推移する.

推移できないとき (表に負の要素を含むとき) は，行と列の選択から再試行する.

一般の MCMC 法の収束性に関しては，経験則は知られているが open である.

5.1 グレブナー基底を用いたマルコフ連鎖構成

定常分布が仮定された分布に従う，既約で非周期的なマルコフ連鎖が構成できれば，分割表をサンプリング出来る.

Diaconis & Sturmfels [15] はマルコフ基底を用いることによって，一様分布へ収束するマルコフ連鎖の代数的構成法を提案した (収束までの時間は示されていない)[15]. 彼らはマルコフ基底が，多項式環のイデアルの生成元集合に等しい事を利用して，適当な項順序 (grevlex 順序,diagnal 項順序等) を用いて被約グレブナー基底を計算してマルコフ基底を得ている.

以下にマルコフ基底を定義する.

定義 2 要素が整数であるような $I \times J$ 分割表，$f_1, \ldots, f_L$ が

1. $f_1, \ldots, f_L$ はすべて，行和，列和が 0 である.
2. 任意の x, x' に対し，

$$x' = x + \sum_{j=1}^{A} \epsilon_j f_{i_j},$$
$$x + \sum_{j=1}^{A} \epsilon_j f_{i_j} \geq 0, \ 1 \leq a \leq A$$

を満たす $(\epsilon_1, f_{i_1}), \ldots, (\epsilon_A, f_{i_A}), \epsilon_i = \pm 1$ が存在する

の 2 条件を満たすとき，$f_1, f_2, \ldots, f_L$ をマルコフ基底と呼ぶ.

すると，以下の定理より，マルコフ基底を用いてマルコフ連鎖を構成することができる.

定理 4 (Hasting [22]) $f_1, \ldots, f_L$ をマルコフ基底，マルコフ連鎖における現在の状態を $x \in \Omega$ とし，$i \in \{1, \ldots, L\}, \epsilon \in \{-1, 1\}$ を各々等確率で独立に選ぶ．

ここで，x から $x + \epsilon f_i$ への推移を

- $x + \epsilon f_i \in \Omega$ ならば，確率 $min\{\frac{H(x+\epsilon f_i)}{H(x)}, 1\}$ で推移する．
- $x + \epsilon f_i \notin \Omega$ ならば，推移しない．

と定めれば，これは Ω 上の，既約，可逆，非周期的なマルコフ連鎖であり，定常分布は多項超幾何分布に比例する． □

上の定理で示したマルコフ連鎖の構成法は一例である．以下の Diaconis & Sturmfels による 2 元分割表の例を紹介する．

例 1 行和 $(1,1,2)$，列和 $(1,1,2)$ の 3×3 分割表全体．*[15]*

1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1
0	1	0	0	0	1	1	0	0	0	1	0	0	1	0	0	1
0	0	2	0	1	1	0	0	2	1	0	1	0	1	1	1	0

例 1 の各分割表を状態とするマルコフ連鎖では，以下のマルコフ基底を加えて推移する．

例 2 例 1 の分割表より得られるイデアルの生成元に対応する表．*[15]*

+1	−1	0	+1	0	−1	0	+1	−1	+1	−1	0	+1	0	−1
−1	+1	0	−1	0	+1	0	−1	+1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	−1	+1	0	−1	0	+1

0	+1	−1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	+1	−1	0	+1	0	−1	0	+1	−1	0	+1	−1
0	−1	+1	−1	+1	0	−1	0	+1	0	−1	+1	0	−1	+1

例 3 推移の例．例 1 の左端の分割表に -1 倍したマルコフ基底を加えると，
左から 3 番目の分割表に推移する．

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} + (-1) \times \begin{bmatrix} +1 & -1 & 0 \\ -1 & +1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

さらに Diaconis & Sturmfels は 2 元分割表について，以下のような結果を示している．

定理 5 (Diaconis & Sturmfels [15]) 周辺和を固定した 2 元分割表全体を頂点，グレブナー基底で移りあえる分割表同士を辺で結んで構成されるグラフ上で，連結なランダムウォークが構成できる． □

定理 6 (Diaconis & Sturmfels [15]) 2 元分割表，輸送問題に対するグレブナー基底が，長さ 4 のサーキット全体となるようなコストベクトルが存在する． □

5.2 グレブナー基底を用いないマルコフ連鎖構成

グレブナー基底を用いたマルコフ連鎖の構成は，グレブナー基底計算に費やす時間が多く，サイズの大きな分割表をサンプリングするためのマルコフ連鎖構成には向いていない．本節ではグレブナー基底を用いないマルコフ連鎖の構成法について紹介する．またマルコフ連鎖は，mixing time と呼ばれる収束するまでの反復回数で，その効率性を評価する．マルコフ連鎖の mixing time は，*total variation* と呼ばれる距離の概念から以下のように定義される [4]．

定義 3 マルコフ連鎖 $\mathcal{M}$ の *mixing time* を以下のように定義する．
$\boldsymbol{\pi} : \Omega \rightarrow [0,1]$: $\mathcal{M}$ 上の定常分布
$\boldsymbol{\pi}' : \Omega \rightarrow [0,1]$: $\mathcal{M}$ 上の任意の確率分布
$\mathrm{D_{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') = (1/2) \sum\{|\boldsymbol{\pi}(x) - \boldsymbol{\pi}'(x)| : x \in \Omega\}$: *total variation*
$\mathrm{P}_x^{(t)}(y)$: $\mathcal{M}$ が初期状態 x から t 回目の推移で y に到達する確率
$\tau_x(\varepsilon) = \min\{t : \forall s \geq t, \mathrm{D_{TV}}(\boldsymbol{\pi}, \mathrm{P}_x^{(s)} \leq \varepsilon)\}$
としたとき，

$$\tau(\varepsilon) = \max\{\tau_x(\varepsilon) : x \in \Omega\}$$

を $\mathcal{M}$ の *mixing time* と呼ぶ．

mixing timeの算定の際には次のような仮定がある．マルコフ連鎖上の相異なる初期状態$X_0, Y_0 (X_0 \neq Y_0)$から出発して，反復して2つの状態の推移を観察する．t回反復を繰り返した結果，同じ状態$X_t (= Y_t)$に到達したら，一方の推移を他方に乗り換える．すなわち，$X_t := Y_t$として以後推移を行うものとする．するとマルコフ連鎖の無記憶性より，$t+1$回以降の反復においては，2つの初期状態X_0, Y_0が異なっていた事を無視できる．この反復を全ての相異なる状態を初期状態として繰り返せば，いつかは一つの推移に収束する．収束した状態は初期状態に依存しないので，マルコフ連鎖の定常分布とみなせる．mixing timeが分割表の行数，列数，$\ln N, \ln(\varepsilon^{-1})$の多項式でおさえられるとき，rapidly mixingであるという．Aldousは，マルコフ連鎖のmixing timeの算定にcoupling法を導入して，mixing timeを算定した．

coupling法は，mixing timeを算定するための道具の一つである．2つの異なる初期状態から出発して，状態が一致するまでの反復回数を計測するための(証明のためだけに用いられる仮想的な)道具である．

定義 4 与えられたマルコフ連鎖$\mathcal{M}$の*coupling*とは以下の条件を満たすものである．ここでΩを$\mathcal{M}$の状態空間の集合とする．

1. 状態空間を$\Omega \times \Omega$とするマルコフ連鎖である．
2. *coupling*の推移確率は次を満たす．
   任意の状態のペア$(x, y) \in \Omega \times \Omega$に対し，

$$\mathrm{P}_M(x, x') = \sum\{\mathrm{P}((x, y), (x', y')) | y' \in \Omega\},$$

$$\mathrm{P}_M(y, y') = \sum\{\mathrm{P}((x, y), (x', y')) | x' \in \Omega\}$$

   が成り立つ．
   ただし，$\mathrm{P}((x, y), (x', y'))$を*coupling*での$(x, y)$から$(x', y')$への推移確率，$\mathrm{P}_M(x, x')$を$\mathcal{M}$における$x$から$x'$への推移確率とする．

coupling法は，$\Omega \times \Omega$上で相異なる2つの状態のペア(X_0, Y_0)が等しい状態となるまでの反復回数を算定する．$\Omega \times \Omega$上の状態のcouplingの仕方はtrivialでは無いが，ここでは省略する．

定理 7 (Coupling lemma, Aldous [3]) $\mathcal{M}$をマルコフ連鎖，X_0, Y_0を$\mathcal{M}$中の任意の異なる2つの初期状態，X_t, Y_tを$\mathcal{M}$中でX_0, Y_0からt回目に推移した状態，εを一様分布

に対する精度 $(0 < \varepsilon < 1)$，$\Pr[X_t \neq Y_t] \leq \varepsilon$ とすると，*mixing time* $\tau(\varepsilon)$ は t でおさえられる．

□

Bubley & Dyer [7] は，mixing time 算定のための道具として，coupling 法を拡張した path coupling 法を提案した．path coupling 法は，2 つの異なる初期状態から定常分布に到達するまでの反復を Ω を頂点とする有向グラフ上で算定する．状態の couple は，有向グラフ上での頂点間の距離を頂点間の有向辺の最短本数で定義し，反復を繰り返すたびに，2 つの状態間の距離が $\beta(< 1)$ 倍の割合で短くなるようにする．couple の決定方法は場合分けが多いので，ここでは割愛するが，Dyer & Greenhill は，path coupling 法を用いて，定常分布が一様分布に収束する，$2 \times J$ 分割表サンプリングのためのマルコフ連鎖の mixing time を算定した．マルコフ連鎖の推移は以下の通りである．

Dyer & Greenhill のマルコフ連鎖の推移方法

Step 1. $2 \times J$ 分割表 x 中の任意の相異なる 2 列 $j_1, j_2 \in J(j_1 < j_2)$ を選ぶ．

Step 2. x' を x の j_1, j_2 列目に $\begin{bmatrix} +\theta & -\theta \\ -\theta & +\theta \end{bmatrix}$ を加えたものとする．

$x' \in Z_{\geq 0}^{2\times J}$ を満たす整数 θ を全て求める．

Step 3. Step 2 で求めた θ を 1 つ等確率で選び，分割表 x' に推移する．

以下に 2×5 分割表の推移例を示す．

例 4 : 2×5 分割表の推移例

$$\begin{bmatrix} 2 & 2 & 4 & 3 & 1 \\ 4 & 1 & 5 & 2 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & \mathbf{2} & 4 & \mathbf{3} & 1 \\ 4 & \mathbf{1} & 5 & \mathbf{2} & 0 \end{bmatrix} + \begin{bmatrix} +\theta & -\theta \\ -\theta & +\theta \end{bmatrix} \quad (\theta = -2, -1, 0, 1)$$

$$1/4 \swarrow \quad 1/4 \downarrow \qquad\qquad 1/4 \downarrow \quad \searrow 1/4$$

$$\begin{bmatrix} 2 & \mathbf{0} & 4 & \mathbf{5} & 1 \\ 4 & \mathbf{3} & 5 & \mathbf{0} & 0 \end{bmatrix} \quad \begin{bmatrix} 2 & \mathbf{1} & 4 & \mathbf{4} & 1 \\ 4 & \mathbf{2} & 5 & \mathbf{1} & 0 \end{bmatrix} \quad \begin{bmatrix} 2 & \mathbf{2} & 4 & \mathbf{3} & 1 \\ 4 & \mathbf{1} & 5 & \mathbf{2} & 0 \end{bmatrix} \quad \begin{bmatrix} 2 & \mathbf{3} & 4 & \mathbf{2} & 1 \\ 4 & \mathbf{0} & 5 & \mathbf{3} & 0 \end{bmatrix}$$

左上の分割表から推移を行うために，ランダムに2列めと4列目を選ぶ．推移可能なθは4つあるので，各々1/4の確率で推移する．ここで$\theta=0$は自分に戻る推移である．

以下に，彼らの提案したマルコフ連鎖のmixing timeを示す．

定理 8 (Path Coupling lemma, Dyer & Greenhill[17]) εを一様分布に対する精度$(0<\varepsilon<1)$とすると，ある正の実数$\beta<1$が存在して，定常分布を一様分布とする，周辺和を固定した$2\times J$分割表をサンプリングするマルコフ連鎖$\mathcal{M}^1$が構築でき，*mixing time* $\tau_1(\varepsilon)$は以下のようにおさえられる．

$$\tau_1(\varepsilon)\leq \ln(D\varepsilon^{-1})/(1-\beta),$$

ただし，Dは*path coupling*法で構築する有向グラフの直径である． □

松井(知), 松井(泰), 小野 [28]は，Dyer & Greenhillの結果を拡張し，定常分布が一様分布に収束する$\underbrace{\{1,2\}\times\{1,2\}\times\cdots\times\{1,2\}}_{m}\times J(=\mathrm{B}^m\times J)$分割表サンプリングのためのマルコフ連鎖を構築した．推移方法はDyer & Greenhillのものを多次元に拡張したものである．以下に$2\times2\times5$分割表の推移例を示す．

例 5 : $2\times2\times5$分割表の推移例

2	2	4	3	1
4	1	5	2	0
1	1	3	1	2
3	2	5	4	2

$\rightarrow$

2	**2**	4	**3**	1
4	**1**	5	**2**	0
1	**1**	3	**1**	2
3	**2**	5	**4**	2

$+$

$+\theta$	$-\theta$
$-\theta$	$+\theta$
$-\theta$	$+\theta$
$+\theta$	$-\theta$

$(\theta=-1,0,1)$

$1/3\swarrow$ $1/3\downarrow$ $\searrow 1/3$

2	**1**	4	**4**	1
4	**2**	5	**1**	0
1	**2**	3	**0**	2
3	**1**	5	**5**	2

2	**2**	4	**3**	1
4	**1**	5	**2**	0
1	**1**	3	**1**	2
3	**2**	5	**4**	2

2	**3**	4	**2**	1
4	**0**	5	**3**	0
1	**0**	3	**2**	2
3	**3**	5	**3**	2

$2\times2\times5$分割表の1段目と2段目を並べて表している．左上の分割表から推移を行うために，ランダムに2列めと4列目を選び，Dyer & Greenhillと同様に等確率で推移を行う．

提案するマルコフ連鎖の mixing time は path coupling 法を用いて算定した．状態の coupling は trivial では無い．

定理 9 (T.Matsui, Y.Matsui & Ono [28]) ε を一様分布に対する精度 $(0 < \varepsilon < 1)$ とすると，定常分布を一様分布とする，周辺和を固定した $\mathrm{B}^m \times J$ 分割表のサンプリングのためのマルコフ連鎖 $\mathcal{M}^2$ が構築でき，*mixing time* $\tau_2(\varepsilon)$ は以下のようにおさえられる．

$$\tau_2(\varepsilon) \leq (1/2)n(n-1)\ln(dn/\varepsilon),$$

ただし，$|J| = n$, d は頻度データ値の平均，すなわち $d = N/(2^m n)$ である．

また，定常分布を多項超幾何分布としても，同じ *mixing time* のマルコフ連鎖が構築できる． □

筆者らが提案したマルコフ連鎖は，mixing time が列数 n，$\ln N, \ln(\varepsilon^{-1})$ の多項式時間でおさえられる rapidly mixing であり，計算の複雑さの観点では効率が良いと見なせる．

例 6 松井 *(知)*，松井 *(泰)*，小野の提案したマルコフ連鎖による *mixing time* 計算例

- $2 \times 2 \times 6$ 分割表，$N = 91$，$\varepsilon = 10^{-10}$ の場合:

$$\begin{aligned} \tau_2(\varepsilon) &\leq (1/2)n(n-1)\ln(dn/\varepsilon) \\ &= (1/2)6(6-1)\ln((91 \times 6)/10^{-10}) \\ &< 356 \end{aligned}$$

- $\mathrm{B}^m \times J$ 分割表 $(|J| = n)$，セル内の平均値 10 程度 $(d = N/(n2^m) = 10)$ とした場合:

$$\begin{aligned} \tau_2(\varepsilon) &\leq (1/2)n(n-1)\ln(dn/\varepsilon) \\ &= (1/2)n(n-1)\ln(10n/\varepsilon) \end{aligned}$$

単位千回

n	5	10	20	40	50	100
$\varepsilon = 10^{-10}$	0.3	1.2	5.3	22	35	148
$\varepsilon = 10^{-20}$	0.3	1.5	6.3	26	42	171

Welsh [35] により提示された，一般の $I \times J$ 分割表に対する rapidly mixing なマルコフ連鎖の構築は open である．例えば，$3 \times J$ 分割表に対し，Dyer & Greenhill のマルコフ連

鎖を適用しても mixing time の算定が出来ない．推移を繰り返す毎に必ず距離が短くなるような coupling を構成できないからである．別の mixing time の算定方法を提案するか，何らかの工夫を施したマルコフ連鎖の構築すれば rapidly mixing なマルコフ連鎖が実現できるかもしれない．

また，多項超幾何分布に従う 2 元分割表は，マルコフ連鎖を構成せずに，$O(N)$ で完全サンプリングできる事が知られている．

6 関連する研究

青木, 竹村 [5] は，$3 \times 3 \times K$ 分割表に対して，極小マルコフ基底を決定し，さらに極小マルコフ基底であるための必要十分条件を一般の離散空間 (分割表を含む) で特徴付けた．坂田, 澤江,Kroumov[31] は，3 元分割表の解析に必要なグレブナー基底を $4 \times 4 \times 4$ のサイズまで計算し，実際に分割表データに適用している．また彼らは分割表の条件付逐次検定を提案し，その性質を調べている．具体的には，2 元分割表の場合，超幾何偏微分方程式における佐々木の creation 作用素が，分割表の集合の生成関数に対して，周辺和を 1 つあげる作用を持つことを利用して，サンプリングを行うごとに逐次検定を構成した．さらに $3 \times 3 \times 3$ 分割表の場合は，6 変数を動かし，青木，竹村の提示した極小マルコフ基底を組み込んだ MCMC 法で p 値を推定しながら逐次検定を行っている．

参考文献

[1] A. AGRESTI, A survey of exact inference for contingency tables, *Statistical Science*, 7 (1992), pp. 131–153.

[2] A. AGRESTI, *Categorical Data Analysis*, John Wiley & Sons, 2002.

[3] D. ALDOUS, Some inequalities for reversible Markov chains, *Journal of the London Mathematical Society*, 25 (1982), pp. 564–576.

[4] D. ALDOUS, Random walks on finite groups and rapidly mixing Markov chains, in A. Dold and B. Eckmann, eds., *Séminaire de Probabilités XVII 1981/1982*, vol. 986 of Springer-Verlag Lecture Notes in Mathematics, Springer-Verlag, New York, (1983), pp. 243–297.

[5] S. AOKI AND A. TAKEMURA, Minimal basis for connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals, *Technical Report METR 2002-02*, Dept. of Mathematical Engineering and Information Physics, Faculty of Engineering, The University of Tokyo, 2002.

[6] D. AVIS AND K. FUKUDA, A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra, *Discrete Comput. Geom.*, 8, (1992), pp. 295–313.

[7] R. BUBLEY AND M. DYER, Path coupling: A technique for proving rapid mixing in Markov chains, *38th Annual Symposium on Foundations of Computer Science*, IEEE, San Alimitos, (1997), pp. 223–231.

[8] R. BUBLEY *Randomized Algorithms : Approximation, Generation, and Counting*, Springer-Verlag, New York, 2001.

[9] F. R. K. CHUNG, R. L. GRAHAM, AND S. T. YAU, On sampling with Markov chains, *Random Structures and Algorithms*, 9 (1996), pp. 55–77.

[10] J. DE LOERA AND S. ONN, The Complexity of Tree-Way Statistical Tables, preprint, 2002.

[11] J. DE LOERA AND B. STURMFELS, Algebraic Unimodular counting, preprint, 2000.

[12] P. DIACONIS AND B. EFFRON, Testing for independence in a two-way table: new interpretations of the chi-square statistics (with discussion), *Annals of Statistics*, 13 (1985), pp. 845-913.

[13] P. DIACONIS AND A. GANGOLLI, Rectangular arrays with fixed margins, in D. ALDOUS, P. P. VARAIYA, J. SPENCER, AND J. M. STEELE (Eds.), *IMA Volumes on Mathematics and its Applications*, 72 (1995), pp. 15–41, Springer, New York.

[14] P. DIACONIS AND L. SALOFF-COSTE, Random walk on contingency tables with fixed row and column sums, Technical Report, Department of Mathematics, (1995), Harvard University.

[15] P. DIACONIS AND B. STRUMFELS, Algebraic algorithms for sampling from conditional distributions, *The Annals of Statistics*, 26 (1998), pp. 363–397.

[16] M. DYER AND A. M. FRIEZE, On the complexity of computing the volume of a polyhedron, *SIAM J. Comput.*, 17 (1988), pp. 27–37.

[17] M. DYER AND C. GREENHILL, A more rapidly mixing Markov chain for graph colourings, *Random Structures and Algorithms*, 13 (1998), pp. 285–317.

[18] M. DYER AND C. GREENHILL, Polynomial-time counting and sampling of two-rowed contingency tables, *Theoretical Computer Sciences*, 246 (2000), pp. 265–278.

[19] M. DYER, R. KANNAN, AND J. MOUNT, Sampling contingency tables, *Random Structures and Algorithms*, 10 (1997), pp. 487–506.

[20] R. A. FISHER, Statistical Methods for Research Workers, *Oliver and Boyde*, Edinburgh, 1934.

[21] M. R. GAREY AND D. S. JOHNSON, Computers and Intractability, A Guide to the Theory of $\mathcal{NP}$-Completeness W. H. Freeman and company, New York, 1979.

[22] W. K. HASTING, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57 (1970), pp. 97–109.

[23] D. HERNEK, Random generation of $2 \times n$ contingency tables, *Random Structures and Algorithms*, 13 (1998), pp. 71–79.

[24] 日比孝之, 可換代数と組合せ論, シュプリンガー・フェアラーク東京 (株),1995.

[25] R. W. IRVING AND M. R. JERRUM, Three dimensional statistical data security problems, *SIAM Journal on Computing*, 23 (1994), pp. 170–184.

[26] M. R. JERRUM AND A. SINCLAIR, The Markov chain Monte Carlo method: an approach to approximate counting and integration, in *Approximation Algorithm for NP-hard problems*, D. S. HOCHBAUM (Ed.), PWS publishing, Boston, 1997, pp. 482–520.

[27] R. KANNAN, P. TETALI AND S. VEMPALA, Simple Markov chain algorithm for generating bipartite graphs and tournaments, in *8th Annual Symposium on Discrete Algorithms*, ACM-SIAM, San Francisco, California, 1997, pp. 193–200.

[28] T. MATSUI, Y. MATSUI AND Y. ONO, Random Generation of $B^m \times J$ Contingency Tables, preprint, 2002.

[29] C. R. MEHTA AND N. R. PATEL, A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables, *Journal of the American Statistical Association*, 78(1983), pp. 427–434.

[30] K. PEARSON, On the χ^2 test of goodness of fit, *Biometrika*, 14 (1922), pp. 186–191.

[31] T.SAKATA, R. SAWAE AND V. KROUMOV, Applications of Gröbner Basis to Analysis of Contingency Tables and Integer Programming, preprint, 2002.

[32] A. SCHRIJVER, Theory of Linear and Integer Programming, Wiley, Chichester, 1986.

[33] B. STURMFELS, Gröbner Bases and Convex Polytopes, University Lecture Notes Series, 8, American Mathematical Society, 1995.

[34] A. TAKEMURA AND S. AOKI, Some characterizations of minimal Markov basis for sampling from discrete conditional distributions *Technical Report METR 2002-04*, Dept. of Mathematical Engineering and Information Physics, Faculty of Engineering, The University of Tokyo, 2002.

[35] D. WELSH, The computational complexity of some classical problems from statistical physics, in *Disorder in Physical Systems*, Oxford University Press, 1990, pp. 307–321.

[36] D. WELSH, Approximate Counting, in *Surveys in Combinatorics*, edited by R.A. Bailey, London Mathematical Society Lecture Notes, 241 (1997), pp. 287–323.